CLEF: Ongoing Activities and Plans for the Future

Maristella Agosti[1] Giorgio Maria Di Nunzio[1] Nicola Ferro[1] Carol Peters[2]

[1]Department of Information Engineering, University of Padua, Italy
{agosti, dinunzio, ferro}@dei.unipd.it

[2]ISTI-CNR, Area di Ricerca – 56124 Pisa – Italy
carol.peters@isti.cnr.it

Abstract

Multilingual information access and cross-language search and retrieval are key issues for digital libraries. For this reason, DELOS has continued to support the activities of the Cross-Language Evaluation Forum (CLEF), which has the promotion of research in multilingual information retrieval system development as its main goal. This paper describes the evolution of CLEF over the last seven years, illustrates the main results, and provides some recommendations for future work in this area.

Keywords: Multilingual Information Access, Cross-Language Information Retrieval, Scientific Data, Data Curation, Long-term Preservation, Test Collections

1 Introduction

The *Cross-Language Evaluation Forum (CLEF)* promotes multilingual system research and development through the organization of annual evaluation campaigns in which a series of tracks designed to test different aspects of mono- and cross-language information retrieval are offered. The intention is to encourage experimentation with all kinds of multilingual information access - from the development of systems for monolingual retrieval operating on many languages to the implementation of complete multilingual multimedia search services. This has been achieved by offering an increasingly complex and varied set of evaluation tracks over the years. The aim is not only to meet but also to anticipate the emerging needs of the R&D community and to encourage the development of next generation multilingual IR systems.

In the following sections, we briefly describe the organization of the CLEF campaigns and (some of) the results achieved. In the final section, we attempt to analyse the actual limits of an evaluation campaign of this type and make some proposals for future directions.

2 History of CLEF Campaigns

CLEF actually began life in 1997 as a track for *Cross Language Information Retrieval (CLIR)* within *Text REtrieval Conference (TREC)*[1], the well-known conference series sponsored in the US by the *National Institute of Standards and Technology (NIST)* and the *Defense Advanced Research Projects Agency (DARPA)*. At that time, almost all existing cross-language systems were designed for text retrieval and handled only two languages, searching from the language of the query to the language of the target collection. In addition, for most of these systems one of the two languages was English.

The development of effective multilingual access functionality is a key issue in the digital library domain. Unfortunately, very few operational digital library systems go much beyond implementing some basic functionality for monolingual search and retrieval in multiple languages, or perhaps some basic cross-language query mechanism using a multilingual thesaurus or simple controlled vocabulary. For this reason, since it began life in 2000 as a Network of Excellence under the Fifth Framework programme of the European Commission, DELOS[2] has supported CLEF.

Thus, when with the encouragement of DELOS the coordination of this activity was moved to Europe and CLEF was launched as an independent initiative, our primary goals were the promotion of system testing and evaluation for European languages other than English and the development of truly multilingual retrieval systems, capable of retrieving relevant information from collections in many languages and in mixed media.

[1]http://trec.nist.gov/
[2]http://www.delos.info/

The CLEF evaluation campaigns have been designed in order to work towards these goals. Tracks are proposed to examine particular areas of cross-language *Information Retrieval (IR)* and are subdivided into tasks, which can vary from year to year, according to the specific aspects of system performance to be tested. Table 1 shows how the number of tracks has been extended since 2000 with the gradual addition of new tracks to reach a total of eight in 2005. CLEF 2006 repeated these same eight tracks but a number of new tasks were introduced within several of the tracks.

3 Experimental Collections

CLEF campaigns adopt a comparative evaluation approach in which system performances are compare according to the Cranfield methodology, which makes use of *experimental collections* [7]. An experimental collection is a triple $\mathcal{C} = (D, T, J)$, where: D is a set of documents, called also collection of documents; T is a set of topics, which expresses the user's information needs and from which the actual queries are derived; J is a set of relevance judgements, i.e. for each topic $t \in T$ and for each document $d \in D$ it is determined whether d is relevant to t or not.

An experimental collection $\mathcal{C}$ allows the comparison of information access systems according to some measurements which quantify their performances. The main goal of an experimental collection is both to provide a common test-bed to be indexed and searched by information access systems and to guarantee the possibility of replicating the experiments. The test collections are extended each year with the addition of new material.

A number of different document collections were used in CLEF 2006 to build the test collections:

- CLEF multilingual comparable corpus of more than 2 million news docs in 12 languages (Bulgarian, Dutch, English, French, Finnish, German, Hungarian, Italian, Portuguese, Russian, Spanish, Swedish). Parts of this collection were used in three 2006 tracks: Ad-Hoc (all languages except Finnish, Swedish and Russian), Question Answering (all languages except Finnish, Hungarian, Swedish and Russian) and GeoCLEF (English, German, Portuguese and Spanish).

- The CLEF domain-specific collection consisting of the GIRT-4 social science database in English and German (over 300,000 documents) and two Russian databases: the Russian Social Science Corpus (approx. 95,000 documents) and the Russian ISISS collection for sociology and economics (approx. 150,000 docs). Controlled vocabularies in German-English and German-Russian were also made available to the participants in this track. This collection was used in the domain-specific track.

- The ImageCLEF track used four collections:
  - the ImageCLEFmed radiological medical database based on a dataset containing images from the Casimage, MIR, PEIR, and PathoPIC datasets (about 50,000 images) with case notes in English (majority) but also German and French;
  - the IRMA collection in English and German of 10,000 images for automatic medical image annotation;
  - the IAPR TC-12 database of 25,000 photographs with captions in English, German and Spanish;
  - a general photographic collection for image annotation provide by LookThatUp (LTUtech) database.

- The Speech retrieval track used the Malach collection of spontaneous conversational speech derived from the Shoah archives in English (more than 750 hours) and Czech (approx 500 hours);

- The WebCLEF track used a collection crawled from European governmental sites, called EuroGOV. This collection consists of more than 3.35 million pages from 27 primary domains. The most frequent languages are Finnish (20%), German (18%), Hungarian (13%), English (10%), and Latvian (9%).

For each collection, appropriate sets of search requests and associated relevance assessments have been built. These test suites form extremely valuable and reusable resources. They are created according to rigorous guidelines and are tested to confirm their stability.

The CLEF Test Suite consisting of the data created for the monolingual, bilingual, multilingual and domain-specific text retrieval tracks for the CLEF 2000-2003 Campaigns is now publicly available. It consists of multilingual document collections in eight languages; step-by-step documentation on how to perform a system evaluation; tools for results computation; multilingual sets of topics; multilingual sets of relevance assessments; guidelines for participants (in English); tables of the results obtained by the participants; publications[3].

[3]The Evaluation Package is now available in the *European Language Resources Association (ELRA)* catalogue (ref. ELRA-E0008). Information can be found at: `http://catalog.elra.info/`

Table 1: CLEF 2000 - 2005: increase in tracks and scope.

CLEF 2000 [tracks]	• mono-, bi- and multilingual textual document retrieval (Ad Hoc) • mono- and cross-language information on structured scientific data (Domain-Specific)
CLEF 2001 [added]	• interactive cross-language retrieval (iCLEF)
CLEF 2002 [added]	• cross-language spoken document retrieval (CL-SR)
CLEF 2003 [added]	• multiple language question answering (QA@CLEF) • cross-language retrieval in image collections (ImageCLEF)
CLEF 2005 [added]	• multilingual retrieval of Web documents (WebCLEF) • cross-language geographical retrieval (GeoCLEF)

4 Results

In this section, we outline some of the principal results achieved by CLEF with respect to the main goal of promoting the development of multilingual information retrieval systems. For complete documentation on individual CLEF experiments and results, track by track and year by year, see the on-line CLEF Working Notes at `http://www.clef-campaign.org/`.

4.1 Cross-language Text Retrieval

CLEF has tried to encourage groups to work their way up gradually from mono- to true multilingual text retrieval by providing them with facilities to test and compare search and access techniques over many languages, pushing them to investigate the issues involved in processing a growing number of languages with different characteristics. As can be seen from Table 1, we have now created ad-hoc cross-language test collections for twelve European languages.

As can be seen from the table, over the years the language combinations have increased and the tasks offered have grown in complexity until, in CLEF 2003, the multilingual track included a task which entailed searching a collection in 8 languages, selected to cover a range of language typologies and linguistic features (Multi-8). We also encouraged system testing with uncommon language pairs (e.g. German to Italian or French to Dutch) in both 2003 and 2004. In 2006, we offered a bilingual task aimed at encouraging system testing with non-European languages against an English target collection. Topics were thus also supplied in Amharic, Oromo, Hindi, Telugu and Indonesian as well as in the usual European languages.

The multilingual task in CLEF 2005 was designed to focus on a particular aspect of the multilingual retrieval problem faced in CLEF 2003: the merging of results over different languages and collections. In CLEF 2006, in the ad-hoc track we included the ”robust” task, a task that emphasizes the importance of stable performance over languages instead of high average performance.

4.1.1 Performance Improvement

Groups submitting results over several years have shown flexibility in advancing to more complex tasks. Much work has been done on fine-tuning for individual languages while other efforts have concentrated on developing language-independent strategies. However, an important question is whether we can demonstrate improvements in system performance. As test collections and tasks vary over years, such improvements are not easy to document. For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. Some findings are reported here.

In 1997, at TREC-6, the best cross-language text retrieval systems had the following results:

Table 2: CLEF 2000 - 2006 Ad-Hoc Tasks

Campaign	Monolingual	Bilingual	Multilingual
CLEF 2000	de; fr; it	x→en	x→de,en,fr,it
CLEF 2001	de; es; fr; it; nl	x→en; x→nl	x→de,en,es,fr,it
CLEF 2002	de; es; fi; fr; it; nl; sv	x→de/es/fi/fr/it/nl/sv; x→en (newcomer)	x→de,en,es,fr,it
CLEF 2003	de; es; fi; fr; it; nl; ru; sv	it→es; de→it; fr→nl; fi→de x→ru; x→en	x→de,en,es,fr; x→de,en,es,fi,fr,it,nl,sv
CLEF 2004	fi; fr; pt; ru	es/fr/it/ru→fi; de/fi/nl/sv→fr; x→ru; x→en	x→fi,fr,pt,ru
CLEF 2005	bg; fr; hu; pt	x→bg/fr/hu/pt	multi8 2 years on multi8 merge
CLEF 2006	bg; fr; hu; pt	x→bg/fr/hu/pt; am/hi/id/te/or→en	robust: x→de,en,es,fr,it,nl

For language abbreviation, we used the ISO-639-1 two-letter codes: am=Amharic; bg=Bulgarian; de=German; en=English; es=Spanish; fi=Finnish; fr=French; hi=Hindi; hu=Hungarian; id=Indonesian; it=Italian; nl=Dutch; or=Oromo; pt=Portuguese; ru=Russian; sv=Swedish.

- en→fr: 49% of best monolingual French system
- en→de: 64% of best monolingual German system

In 2002, at CLEF, where there was no restriction on topic and target language, the best systems gave:

- en→fr: 83,4% of best monolingual French system
- en→de: 85,6% of best monolingual German system

CLEF 2003 enforced the use of “unusual” language pairs, with the following impressive results:

- it→es: 83% of best monolingual Spanish IR system
- de→it: 87% of best monolingual Italian IR system
- fr→nl: 82% of best monolingual Dutch IR system

In CLEF 2005, where we introduced two new languages, we found:

- x→fr: 85% of best monolingual French system
- x→pt: 88% of best monolingual Portuguese system
- x→bg: 74% of best monolingual Bulgarian system
- x→hu: 73% of best monolingual Hungarian system

In CLEF 2006, for the same target languages, we had the we had the following results:

- x→fr: 94% of best monolingual French system
- x→pt: 91% of best monolingual Portuguese system
- x→bg: 52.5% of best monolingual Bulgarian system
- x→hu: 51% of best monolingual Hungarian system

We find that with languages for which testing has gone on for several years there is usually little variation in performance between the best groups with the best results close to monolingual performance, whereas for ”new” languages where there has been little CLIR system testing, there is normally room for improvement. It should be noted that the results for Bulgarian and Hungarian in 2006 are not significant. Only one group submitted runs in these languages for the bilingual tasks.

As stated, in CLEF 2005 we attempted to reuse the Multi-8 test collection created in CLEF 2003 to see whether a similar improvement in multilingual system performance could be measured, and also to examine the results merging problem. Unfortunately, there was not a large participation in this task and the results obtained are only indicative. However, we can report that the top performing submissions to both the multilingual 2-Years-On and the merging tasks improved on the performance of the best submission to the CLEF 2003 Multi-8 task.

Summing up, we find that, over the years, CLEF participants learn from each other and build up a collective knowhow. Thus, as time passes, we see a convergence of techniques and results with very little statistical difference between the best systems. We have observed that the best systems are a result of careful tuning of every component, and of combining different algorithms and information sources for every subtask [6].

4.2 Cross-language Information Extraction

For many years, IR has concentrated on document retrieval. However, users often want specific answers rather than all the information that is to be found on a given topic. For this reason, information extraction systems have been given much attention. In 2003, CLEF introduced a cross-language question-answering track thus stimulating the development of some of the very first multilingual *Question Answering (QA)* systems. CLEF 2005 and 2006 ran experiments in cross-language geographic IR.

4.2.1 Multilingual Question Answering

Question answering systems have been evaluated for many years at TREC and the track evolved over the years to offer increasingly difficult tasks. However, multilinguality had never been taken into consideration. As QA techniques are mainly based on natural language processing tools and resources, we felt that it was important to fill this gap in CLEF. The aim of the track is to encourage testing on languages other than English, to check and/or improve the portability of technologies implemented in English QA systems, and to force the QA community to design real multilingual systems. The QA@CLEF campaign in 2006 was the result of experience acquired during the two previous years and proved very popular. The main tasks assessed mono- and bilingual system retrieval for eight target collections. The participating systems were fed a set of 200 questions, which could be about: facts or events (Factoid questions); definitions of people, things or organisations (Definition questions); of people, objects or data (List questions).

Two pilot tasks were also run: the WiQA[4] and AVE[5]. The purpose of the WiQA pilot was to see how IR and *Natural Language Processing (NLP)* techniques can be effectively used to help readers and authors of Wikipedia pages access information spread throughout Wikipedia rather than stored locally on the pages. The Answer Validation Exercise (AVE) encouraged validating the correctness of the answers given by a QA system. The basic idea is that once a pair [answer + snippet] is returned by a QA system, a hypothesis is built by turning the pair [question + answer] into the affirmative form. If the related text (a snippet or a document) semantically entails this hypothesis, then the answer is expected to be correct.

In addition to the tasks proposed during the actual competition, a "time constrained" QA exercise was run by the University of Alicante during the CLEF 2006 Workshop. In order to evaluate the ability of QA systems to retrieve answers in real time, the participants were given a time limit (e.g. one or two hours) in which to answer a set of questions. These question sets are different and smaller than those provided in the main task questions). The initiative is aimed towards providing a more realistic scenario for a QA exercise.

In these four years, performance for both mono- and cross-language question answering systems has shown improvement, with the best non-English systems obtaining very similar results to those of TREC, and the best bilingual systems obtaining a performance of approximately 60% of monolingual results. From a comparison of approaches, we see that most systems pre-process the document collection, adopting linguistic processors and language resources such as *Part of Speech (PoS)* taggers, named entity recognizers, WordNet, gazetteers. Many systems adopt a deep parsing strategy while only a few use any logical representation [16].

4.2.2 Cross-language Geographic Retrieval

After being a pilot track in 2005, GeoCLEF advanced to be a regular track within CLEF 2006. The purpose of GeoCLEF is to test and evaluate cross-language *Geografic Information Retrieval (GIR)*: retrieval for topics with a geographic specification. For GeoCLEF 2006, twenty-five search topics were defined by the organizing groups for searching English, German, Portuguese and Spanish document collections. Topics were translated into English, German, Portuguese, Spanish and Japanese. Several topics in 2006 were significantly more geographically challenging than in 2005. Seventeen groups submitted 149 runs (up from eleven groups and 117 runs in GeoCLEF 2005). The groups used a variety of approaches, including geographic bounding boxes, named entity extraction and external knowledge bases (geographic thesauri and ontologies and gazetteers). The test collection developed for GeoCLEF is the first GIR test collection available to the GIR research community [13].

4.3 Cross-language Multimedia Retrieval

The current growth of multilingual digital material in a combination of different media (e.g. image, speech, video) means that there is an increasing interest in systems capable of automatically accessing the information available in these archives. For this

[4] http://ilps.science.uva.nl/WiQA/
[5] http://nlp.uned.es/QA/AVE/

reason, CLEF supported a preliminary investigation aimed at evaluating systems for cross-language spoken document retrieval in 2002 and in 2003 introduced a track for cross-language retrieval on image collections.

4.3.1 Cross-Language Speech Retrieval

The Cross-Language Speech Retrieval (CL-SR) in CLEF 2003 and 2004 experimented with cross-language retrieval on transcripts of English broadcast news. In 2005 and 2006 the track has focused on spontaneous speech retrieval over languages. Spontaneous speech is considerably more challenging for the *Automatic Speech Recognition (ASR)* techniques on which fully-automatic content-based search systems are based. Recent advances in ASR have made it possible to contemplate the design of systems that would provide a useful degree of support for searching large collections of spontaneous conversational speech, but no representative test collection that could be used to support the development of such systems has been widely available for research use. The principal goal of the CLEF 2005 CL-SR track was thus to create such a test collection. Additional goals included benchmarking the present state of the art for ranked retrieval of spontaneous conversational speech and fostering interaction among a community of researchers with interest in that challenge. The collection used was a set of interviews in English with Holocaust survivors, extracted from the Shoah archives.

A reusable test collection for searching spontaneous conversational English speech using queries in five languages (Czech, English, French, German and Spanish) was built and includes speech recognition for spoken words, manually and automatically assigned controlled vocabulary descriptors for concepts, dates and locations, manually assigned person names, and handwritten segment summaries.

The CL-SR 2006 track included two tasks: to identify topically coherent segments of English interviews in a known-boundary condition, and to identify time stamps marking the beginning of topically relevant passages in Czech interviews in an unknown-boundary condition. Five teams participated in the English evaluation, performing both monolingual and cross-language searches of ASR transcripts, automatically generated metadata, and manually generated metadata. Results indicate that the 2006 evaluation topics were more challenging than those used in 2005, but that cross-language searching continued to pose no unusual challenges when compared with collections of character-coded text. Three teams participated in the Czech evaluation, but no team achieved results comparable to those obtained with English interviews. The reasons for this outcome are not yet clear [20].

4.3.2 ImageCLEF

The ImageCLEF retrieval benchmark aims at evaluating image retrieval from multilingual document collections. Images by their very nature are language independent, but are often accompanied by semantically related texts (e.g. captions or metadata). Images can then be retrieved using primitive features based on pixels which form the contents of an image (e.g. using a visual exemplar), abstracted features expressed through text, or a combination of both. The language used to express the associated texts or textual queries should not affect retrieval, i.e. an image with a caption written in English should be searchable in languages other than English.

A major goal of ImageCLEF is to investigate the effectiveness of combining text and image for retrieval and to promote the exchange of ideas which may help improve the performance of future image retrieval systems. Participants are provided with image collections, representative search requests (expressed by both image and text) and relevance judgements indicating which images are relevant to each search request. ImageCLEF began in 2003 with a first collection of historical photographs with attached metadata in English. In 2004, a domain-specific collection of medical radiographic images with casenotes in French and German was added; this proved to be of great interest to many groups as it represented a real-world application. The medical collection has been considerably expanded over the years.

ImageCLEF 2006 was divided into two main sections (ImageCLEFphoto and ImageCLEFmed) regarding retrieval on colour travel photos and on medical images, respectively. Realistic (and different) scenarios in which to test the performance of image retrieval systems and present different challenges and problems were offered to participants. Both sections included general retrieval and object annotation tasks. Figure 1 shows a sample document from the ImageCLEFphoto collection, while Figure 2 illustrates a document search requests in the ImageCLEFmed task. In both tasks, the best results were obtained by systems that combined text and content-based retrieval mechanisms.

ImageCLEF is important because more research into multimodal retrieval, combining text and visual features and catering also for multilinguality is needed. For this reason, it is not surprising that this track has been very popular in both CLEF 2005 and 2006 with a large participation: twenty five groups in 2006 [8, 17].

4.3.3 Interactive CLEF

In CLEF 2006, the interactive track joined forces with the image track to work on a new type of interactive image retrieval task to better capture the interplay be-

tween image and the multilingual reality of the internet for the public at large. The task was based on the popular image perusal community Flikr , a dynamic and rapidly changing database of images with textual comments, captions, and titles in many languages and annotated by image creators and viewers cooperatively in a self-organizing ontology of tags (a so-called "folksonomy"). Participants built a multilingual search front-end to Flikr and studied the behaviour of users for a given set of searching tasks. the emphasis was put on studying the process rather than evaluating the outcome [14].

5 The Technical Infrastructure as a Scientific Digital Library

Since CLEF 2005, we have adopted a new approach to the design and development of the technical infrastructure which supports the course of the evaluation campaigns.

5.1 Motivations and Objectives

If we consider the Cranfield evaluation methodology and the achievements and outcomes of the evaluation campaigns, it is clear that we deal with different kinds of valuable *scientific data*. Indeed, the experimental collections and the experiments represent our primary scientific data and the starting point of our investigation. Using the experimental data, we produce different performance measurements, such as precision and recall, in order to evaluate the performances of an *Information Retrieval System (IRS)* for a given experiment. Starting from these performance measurements, we compute descriptive statistics, such as mean or median, used to summarize the overall performances achieved by an experiment or by a collection of experiments. Finally, we perform hypothesis tests and other statistical analyses to conduct an in-depth analysis and comparison over a set of experiments.

When we deal with scientific data, "the lineage (provenance) of the data must be tracked, since a scientist needs to know where the data came from [...] and what cleaning, rescaling, or modelling was done to arrive at the data to be interpreted" [1]. Moreover, [15] points out how provenance is "important in judging the quality and applicability of information for a given use and for determining when changes at sources require revising derived information". Furthermore, when scientific data are maintained for further and future use, they are frequently enriched and, sometimes, the enrichment of a portion of scientific data can make use of a *citation* [2, 3]. Finally, [19] highlights that "digital data collections enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration". On the other hand, the Cranfield methodology was developed to create comparable experiments and evaluate the performances of an IRS rather than modeling, managing, and curating the scientific data produced during an evaluation campaign and thus, we need to extend it in order to keep these new factors into account [4, 5]

The growing interest in the proper management of scientific data has been brought to general attention by different world organizations, among them the European Commission, the US National Scientific Board, and the Australian Working Group on Data for Science. The EC in the i2010 Digital Library Initiative clearly states that "digital repositories of scientific information are essential elements to build European eInfrastructure for knowledge sharing and transfer, feeding the cycles of scientific research and innovation up-take" [12]. The US National Scientific Board points out that "organizations make choices on behalf of the current and future user community on issues such as collection access; collection structure; technical standards and processes for data curation; ontology development; annotation; and peer review". And, those organizations "are uniquely positioned to take leadership roles in developing a comprehensive strategy for long-lived digital data collections" [19]. The Australian Working Group on Data for Science suggests to "establish a nationally supported long-term strategic framework for scientific data management, including guiding principles, policies, best practices and infrastructure", that "standards and standards-based technologies be adopted and that their use be widely promoted to ensure interoperability between data, metadata, and data management systems", and that "the principle of open equitable access to publicly-funded scientific data be adopted wherever possible [...] As part of this strategy, and to enable current and future data and information resources to be shared, mechanisms to enable the discovery of, and access to, data and information resources must be encouraged" [21].

Scientific data, their enrichment and interpretation are essential components of scientific research. The Cranfield methodology traces out how these scientific data have to be produced, while the statistical analysis of experiments provide the means for further elaborating and interpreting the experimental results. Nevertheless, the current methodologies does not require any particular coordination or synchronization between the basic scientific data and the analyses on them, which are treated as almost separated items. On the contrary, researchers would greatly benefit from an integrated vision of them, where the access to a scientific data item could also offer the possibility of retrieving all the analyses and interpretations on it. Furthermore, it should be possible to enrich the basic scientific data in an incremental way, progressively adding

further analyses and interpretations on them.

As a consequence, an evaluation campaign has to provide a software infrastructure suitable for carrying out this second new role. In this context, *Digital Library Systems (DLSs)* represent the natural choice for managing, making accessible, citing, curating, enriching, and preserving all the information resource produced during an evaluation campaign. Indeed, [15] points out how *information enrichment* should be one of the activities supported by a DLS and, among the different kinds of it, considers provenance as "important in judging the quality and applicability of information for a given use and for determining when changes at sources require revising derived information". In addition, [15] observes that also *citation*, intended as the possibility of explicitly mentioning and making references to portions of a given digital object, should be part of the information enrichment strategies supported by a DLS.

In conclusion, DLSs can act as the systems of choice to support evaluation campaigns in making a step forward; they are able to both address the key points highlighted above and provide a more mature way of dealing with the scientific data produced during the IR experimental evaluation.

5.2 Architecture

The design of the scientific DLS is built around five main areas of modelling:

- **evaluation campaign**: deals with the different aspects of an evaluation forum, such as the conducted evaluation campaigns and the different editions of each campaign, the tracks along which the campaign is organized, the subscription of the participants to the tracks, the topics of each track;
- **collection**: concerns the different collections made available by an evaluation forum; each collection can be organized into various files and each file may contain one or more multimedia documents; the same collection can be used by different tracks and by different editions of the evaluation campaign;
- **experiments**: regards the experiments submitted by the participants and the evaluation metrics computed on those experiments, such as precision and recall;
- **pool/relevance assessment**: is about the pooling method where a set of experiments is pooled and the documents retrieved in those experiments are assessed with respect to the topics of the track the experiments belongs to;
- **statistical analysis**: models the different aspects concerning the statistical analysis of the experimental results, such as the type of statistical test employed, its parameters, the observed test statistic, and so forth.

Figure 1 shows the architecture of the proposed DLS. It consists of three layers – data, application and interface logic layers – in order to achieve a better modularity and to properly describe the behavior of the service by isolating specific functionalities at the proper layer. In this way, the behavior of the system is designed in a modular and extensible way. In the following, we briefly describe the architecture shown in Figure 1, from bottom to top.

5.2.1 Data Logic

The data logic layer deals with the persistence of the different information objects coming from the upper layers. There is a set of "storage managers" dedicated to storing the submitted experiments, the relevance assessments and so on. We adopt the *Data Access Object (DAO)*[6] and the *Transfer Object (TO)*[6] design patterns. The DAO implements the access mechanism required to work with the underlying data source, acting as an adapter between the upper layers and the data source. If the underlying data source implementation changes, this pattern allows the DAO to adapt to different storage schemes without affecting the upper layers.

In addition to the other storage managers, there is the *log storage manager* which keeps track of both system and user events. It captures information such as the user name, the *Internet Protocol (IP)* address of the connecting host, the action that has been invoked by the user, the messages exchanged among the components of the system in order to carry out the requested action, any error condition, and so on. Thus, besides offering us a log of the system and user activities, the log storage manager allows us to trace the provenance of each piece of data from its entrance in the system to every further processing on it.

Finally, on top of the various "storage managers" there is the *Storing Abstraction Layer (SAL)* which hides the details about the storage management to the upper layers. In this way, the addition of a new "storage manager" is totally transparent for the upper layers.

5.2.2 Application Logic

The application logic layer deals with the flow of operations within the DLS. It provides a set of tools capable of managing high-level tasks, such as experiment submission, pool assessment, statistical analysis of an experiment.

[6] http://java.sun.com/blueprints/corej2eepatterns/Patterns/

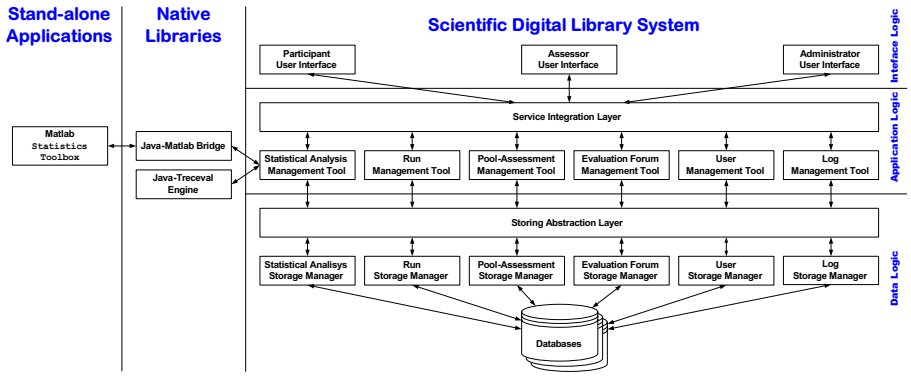

Figure 1: Service architecture for supporting evaluation of information access components.

For example, the *Statistical Analysis Management Tool (SAMT)* offers the functionalities needed to conduct a statistical analysis on a set of experiments. In order to ensure comparability and reliability, the SAMT makes uses of well-known and widely used tools to implement the statistical tests, so that everyone can replicate the same test, even if he has no access to the service. In the architecture, the MATLAB Statistics Toolbox[7] has been adopted, since MATLAB is a leader application in the field of numerical analysis which employs state-of-the-art algorithms, but other software could have been used as well. In the case of MATLAB, an additional library is needed to allow our service to access MATLAB in a programmatic way; other softwares could require different solutions. As an additional example aimed at wide comparability and acceptance of the tools, a further library provides an interface for our service towards the `trec_eval` package[8]. `trec_eval` has been firstly developed and adopted by TREC and represents the standard tool for computing the basic performance figures, such as precision and recall.

Finally, the *Service Integration Layer (SIL)* provides the interface logic layer with a uniform and integrated access to the various tools. As we noticed in the case of the SAL, thanks to the SIL the addition of new tools is transparent for the interface logic layer.

5.2.3 Interface Logic

This is the highest level of the architecture and is the access point for the user to interact with the system. It provides specialised *User Interfaces (UIs)* for different types of users: the participants, the assessors, and the administrators. Note that, thanks to the abstraction provided by the application logic layer, different kind of UIs can be provided, either stand-alone applications or Web-based applications.

[7] `http://www.mathworks.com/products/statistics/`

[8] `ftp://ftp.cs.cornell.edu/pub/smart/`

5.3 Running System

The proposed software infrastructure has been implemented in a prototype, called *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* [9, 11], and has been tested in the context of the CLEF 2005 and 2006 evaluation campaigns. The prototype provides support for:

- the management of an evaluation forum: the track set-up, the harvesting of documents, the management of the subscription of participants to tracks;
- the management of submission of experiments, the collection of metadata about experiments, and their validation;
- the creation of document pools and the management of relevance assessment;
- common statistical analysis tools for both organizers and participants in order to allow the comparison of the experiments;
- common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses;
- common *eXtensible Markup Language (XML)* format for exchanging data between organizers and participants.

DIRECT was successfully adopted during the CLEF 2005 campaign. It was used by nearly 30 participants spread over 15 different nations, who submitted more than 530 experiments. 15 assessors then assessed more than 160,000 documents in seven different languages, including Russian and Bulgarian which use the Cyrillic rather than the Latin alphabet. During the CLEF 2006 campaign, DIRECT was used by nearly 75 participants spread over 25 different nations, who have submitted around 570 experiments. 40 assessors assessed more than 198,500 documents in nine

Figure 2: Participant user interface for the management of the experiments.

different languages. DIRECT was then used to produce reports and overview graphs about the submitted experiments [10].

Figure 2 shows the user interface for the management of the experiments submitted by the participant.

Figure 3 shows the user interface offered to the assessor for making the relevance assessments.

Finally, Figure 4 shows some of the performance measurements and descriptive statistics available to the participants.

6 Conclusions

The results achieved by CLEF in these years are impressive. We can summarise them in the following main points:

- implementation of a powerful and flexible technical infrastructure including data curation functionality;
- creation of important, reusable test collections for system benchmarking;
- building of a strong, multidisciplinary research community;
- R&D activity in new areas such as cross-language question answering, multilingual retrieval for mixed media, and cross-language geographic information retrieval;
- documented improvement in system performance for cross-language text retrieval systems.

Furthermore, CLEF evaluations have provided qualitative and quantitative evidence along the years as to which methods give the best results in certain key areas, such as multilingual indexing, query translation, resolution of translation ambiguity, results merging.

However, although CLEF has done much to promote the development of multilingual IR systems, so far the focus has been on building and testing research prototypes rather than developing fully operational systems. There is still a considerable gap between the research and the application communities and, despite the strong demand for and interest in multilingual IR functionality, there are still very few commercially viable systems on offer. The challenge that CLEF must face in the near future is how to best transfer the research results to the market place. CLEF 2006 took a first step in this direction with the organization of the real time exercise as part of the question-answering track. This experiment will be repeated in CLEF 2007. New metrics will be introduced into the ad-hoc track in order to favour systems that achieve a high precision of correct responses in the first ten results returned - rather than a good average precision. This is a user-oriented measure and makes more sense in the internet dominated world. We also intend to focus more on the multilingual web searching tasks in the future. Content on the world wide web is essentially multilingual, and web users are often polyglots. In addition, the interactive track will be extended and more attention will be given to aspects involving user satisfaction. issues.

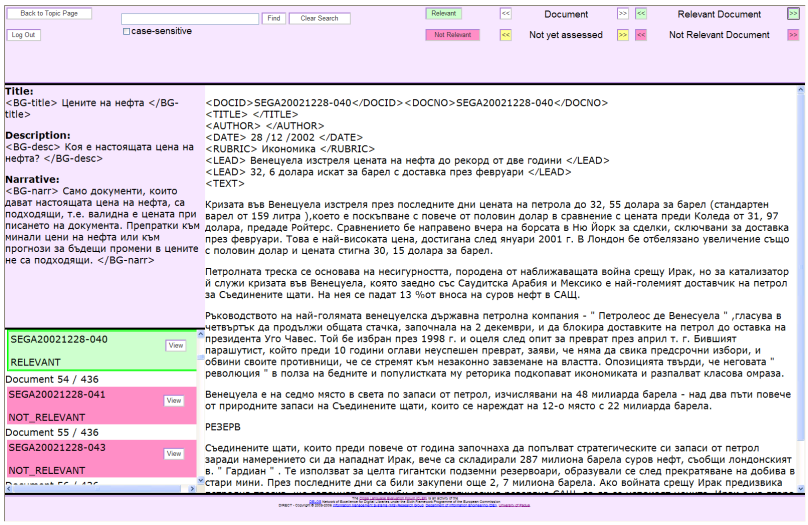

Figure 3: Assessor user interface for performing the relevance assessments.

However, all this is not sufficient. In our opinion, if the gap between academic excellence and commercial adoption of CLIR technology is to be bridged, we need to extend the current CLEF formula in order to give application communities the possibility to benefit from the CLEF evaluation infrastructure without the need to participate in academic exercises that may be irrelevant to their current needs. We feel that CLEF should introduce an application support structure aimed at encouraging take-up of the technologies tested and optimized within the context of the evaluation exercises. This structure would provide tools, resources, guidelines and consulting services to applications or industries that need to include multilingual functionality within a service or product.

Acknowledgements

The work reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

References

[1] S. Abiteboul, R. Agrawal, P. Bernstein, M. Carey, S. Ceri, B. Croft, D. DeWitt, M. Franklin, H. Garcia-Molina, D. Gawlick, J. Gray, L. Haas, A. Halevy, J. Hellerstein, Y. Ioannidis, M. Kersten, M. Pazzani, M. Lesk, D. Maier, J. Naughton, H.-J. Schek, T. Sellis, A. Silberschatz, M. Stonebraker, R. Snodgrass, J. D. Ullman, G. Weikum, J. Widom, and S. Zdonik. The Lowell Database Research Self-Assessment. *Communications of the ACM (CACM)*, 48(5):111–118, 2005.

[2] M. Agosti, G. M. Di Nunzio, and N. Ferro. A Data Curation Approach to Support In-depth Evaluation Studies. In F. C. Gey, N. Kando, C. Peters, and C.-Y. Lin, editors, *Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006)*, pages 65–68. `http://ucdata.berkeley.edu/sigir2006-mlia.htm` [last visited 2007, March 23], 2006.

[3] M. Agosti, G. M. Di Nunzio, and N. Ferro. Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In Nardi et al. [18].

[4] M. Agosti, G. M. Di Nunzio, and N. Ferro. A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns. In T. Sakai and M. Sanderson, editors, *The First International Workshop on Evaluating Information Access (EVIA 2007)*. (in print), May 2007.

[5] M. Agosti, G. M. Di Nunzio, and N. Ferro. The Importance of Scientific Data Curation for Evaluation Campaigns. In C. Thanos and F. Borri, editors, *DELOS Conference 2007 Working Notes*, pages 185–193. ISTI-CNR, Gruppo ALI, Pisa, Italy, February 2007.

[6] M. Braschler and C. Peters. Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval*, 7(1–2):7–31, 2004.

[7] C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In K. Spack Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, California, USA, 1997.

[8] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the ImageCLEF 2006 Photographic Retrieval and Object Annotation Tasks. In Nardi et al. [18].

[9] G. M. Di Nunzio and N. Ferro. DIRECT: a System for Evaluating Information Access Components of Digital Libraries. In A. Rauber, S. Christodoulakis, and A. Min Tjoa, editors, *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, pages 483–484. Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 2005.

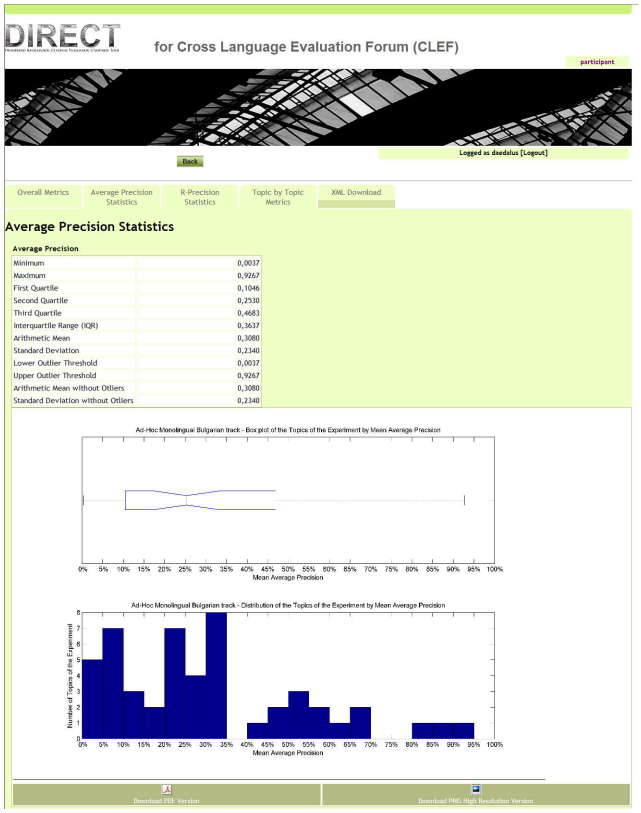

Figure 4: Performance measurements and descriptive statistics available to participants.

[10] G. M. Di Nunzio and N. Ferro. Appendix A: Results of the Ad-hoc Bilingual and Monolingual Tasks. In Nardi et al. [18].

[11] G. M. Di Nunzio and N. Ferro. Scientific Evaluation of a DLMS: a service for evaluating information access components. In J. Gonzalo, C. Thanos, M. F. Verdejo, and R. C. Carrasco, editors, *Proc. 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006)*, pages 536–539. Lecture Notes in Computer Science (LNCS) 4172, Springer, Heidelberg, Germany, 2006.

[12] European Commission Information Society and Media. i2010: Digital Libraries. `http://europa.eu.int/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf` [last visited 2007, March 23], October 2006.

[13] F. Gey, R. Larson, M. Sanderson, K. Bischoff, T. Mandl, K. Womser-Hacker, D. Santos, P. Rocha, G. M. Di Nunzio, and N. Ferro. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In Nardi et al. [18].

[14] J. Gonzalo, J. Karlgren, and P. Clough. iCLEF 2006 Overview: Searching the Flickr WWW Photo-Sharing Repository. In Nardi et al. [18].

[15] Y. Ioannidis, D. Maier, S. Abiteboul, P. Buneman, S. Davidson, E. A. Fox, A. Halevy, C. Knoblock, F. Rabitti, H.-J. Schek, and G. Weikum. Digital library information-technology infrastructures. *International Journal on Digital Libraries*, 5(4):266–274, 2005.

[16] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, P. Osenova, A. Peñas, V. Jijkoun, B. Sacaleanu, P. Rocha, and R. Sutcliffe. Overview of the CLEF 2006 Multilingual Question Answering Track. In Nardi et al. [18].

[17] H. Müller, T. Deselaers, T. Lehmann, P. Clough, E. Kim, and W. Hersh. Overview of the ImageCLEFmed 2006 Medical Retrieval and Annotation Tasks. In Nardi et al. [18].

[18] A. Nardi, C. Peters, and J. L. Vicedo, editors. *Working Notes for the CLEF 2006 Workshop*. Published Online, 2006.

[19] National Science Board. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century (NSB-05-40)*. National Science Foundation (NSF). `http://www.nsf.gov/pubs/2005/nsb0540/` [last visited 2007, March 23], September 2005.

[20] D. W. Oard, J. Wang, G. J. F. Jones, R. W. White, P. Pecina, D. Soergel, X. Huang, and I. Shafran. Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In Nardi et al. [18].

[21] Working Group on Data for Science. *FROM DATA TO WISDOM: Pathways to Successful Data Management for Australian Science*. Report to Ministers Science, Engineering and Innovation Council (PMSEIC) , `http://www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/Presentation_Data_for_Science.htm` [last visited 2007, March 23], December 2006.